Generating counterfactual explanations of tumor spatial proteomes to discover therapeutic strategies for enhancing immune infiltration

Zitong Jerry Wang[1*], Alexander M. Xu[2], Aman Bhargava[1] and Matt W. Thomson[1*]

[1]Division of Biology and Biological Engineering, California Institute of Technology, 1200 E California Blvd, Pasadena, 91125, CA, USA.

[2]Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, 8700 Beverly Blvd, Los Angeles, 90048, CA, USA.

*Corresponding author(s). E-mail(s): zwang2@caltech.edu; mthomson@caltech.edu;

Abstract

Immunotherapies can halt or slow down cancer progression by activating either endogenous or engineered T cells to detect and kill cancer cells. For immunotherapies to be effective, T cells must be able to infiltrate the tumor microenvironment. However, many solid tumors resist T-cell infiltration, challenging the efficacy of current therapies. Here, we introduce Morpheus, an integrated deep learning framework that takes large scale spatial omics profiles of patient tumors, and combines a formulation of T-cell infiltration prediction as a self-supervised machine learning problem with a counterfactual optimization strategy to generate minimal tumor perturbations predicted to boost T-cell infiltration. We applied our framework to 368 metastatic melanoma and colorectal cancer (with liver metastases) samples assayed using 40-plex imaging mass cytometry, discovering cohort-dependent, combinatorial perturbations, involving CXCL9, CXCL10, CCL22 and CCL18 for melanoma and CXCR4, PD-1, PD-L1 and CYR61 for colorectal cancer, predicted to support T-cell infiltration across large patient

cohorts. Our work presents a paradigm for counterfactual-based prediction and design of cancer therapeutics using spatial omics data.

Introduction

The immune composition of the tumor microenvironment (TME) plays a crucial role in determining patient prognosis and response to cancer immunotherapies [1–3]. Immunotherapies that alter the immune composition using transplanted or engineered immune cells (chimeric antigen receptor T cell therapy) or remove immunosuppressive signaling (checkpoint inhibitors) have shown exciting results in relapsed and refractory tumors in hematological cancers and some solid tumors. However, effective therapeutic strategies for most solid tumors remain limited [4–6]. The TME is a complex mixture of immune cells, including T cells, B cells, natural killer cells, and macrophages, as well as stromal cells and tumor cells [1]. The interactions between these cells can either promote or suppress tumor growth and progression, and ultimately impact patient outcomes. For example, high levels of tumor-infiltrating lymphocytes (TILs) in the TME are associated with improved prognosis and response to immunotherapy across multiple cancer types [7, 8]. Conversely, an immunosuppressive TME characterized by low levels of TILs is associated with poor prognosis and reduced response to immunotherapy [9]. Durable, long-term clinical response of T-cell-based immunotherapies are often constrained by a lack of T-cell infiltration into the tumor, as seen in classically "cold" tumors such as triple-negative breast cancer or pancreatic cancer, which have seen little benefit from immunotherapy [10–12]. The precise cellular and molecular factors that limit T-cell infiltration into tumors is an open question.

Spatial omics technologies capture the spatial organization of cells and molecular signals in intact human tumors with unprecedented molecular detail, revealing the relationship between localization of different cell types and tens to thousands of molecular signals [13]. T-cell infiltration is modulated by a rich array of signals within the tumor microenvironment (TME) such as chemokines, adhesion molecules, tumor antigens, immune checkpoints, and their cognate receptors [14]. Recent advances in *in situ* molecular profiling techniques, including spatial transcriptomic [15, 16] and proteomic [17, 18] methods, simultaneously capture the spatial relationship of tens to thousands of molecular signals and T cell localization in intact human tumors with micron-scale resolution. Imaging mass cytometry (IMC) is one such technology that uses metal-labeled antibodies to enable simultaneous detection of up to 40 antigens and transcripts in intact tissue [17].

Recent work on computational methods as applied to multiplexed tumor images have primarily focused on predicting patient-level phenotypes such as survival, by identifying spatial motifs from tumor microenvironments [19–22].

These methods have generated valuable insights into how the complex composition of TMEs influences patient prognosis and treatment response, but they fall short of generating concrete, testable hypotheses for therapeutic interventions that may improve patient outcomes. Given the prognostic significance of T-cell infiltration into tumors, we need computational tools that can predict immune cell localization from environmental signals and systematically generate specific, feasible tumor perturbations that are predicted to alter the TME to improve patient outcomes.

Counterfactual explanations (CFEs) can provide important insight in image analysis applications [23], but have not been applied to multiplexed imaging data. Traditionally, CFEs help clarify machine learning model decisions by exploring hypothetical scenarios, showing how the model's interpretation would change if a feature in an image were altered slightly [24]. For instance, slight pixel intensity variations or minor edge alterations in a tumor's appearance on an X-ray might lead a diagnostic model to classify the scan differently. Numerous CFE algorithms exist to elucidate a model's decision boundaries and shed light on its sensitivity to specific image features [25]. In multiplexed tissue images where each pixel captures detailed molecular information, variations in pixel intensity directly correspond to specific molecular interventions. Thus, spatial omics data enables the extension of CFEs from understanding to predicting actionable interventions.

In this work, we introduce Morpheus, an integrated deep learning framework that first leverages large scale spatial omics profiles of patient tumors to formulate T-cell infiltration prediction as a self-supervised machine learning (ML) problem, and combines this prediction task with counterfactual optimization to propose tumor perturbations that are predicted to boost T-cell infiltration. Specifically, we train a convolutional neural network to predict T-cell infiltration using spatial maps of the TME provided by IMC. We then apply a gradient-based counterfactual generation strategy to the infiltration neural network to compute changes to the signaling molecule levels that increase predicted T-cell abundance. We apply Morpheus to melanoma [26] and colorectal cancer (CRC) with liver metastases [27] to discover tumor perturbations that are predicted to support T cell infiltration in tens to hundreds of patients. We provide further validation of ML-based T-cell infiltration prediction using an additional breast cancer data set [28]. For patients with melanoma, Morpheus predicts combinatorial perturbation to the CXCL9, CXCL10, CCL22 and CCL18 levels can convert immune-excluded tumors to immune-inflamed in a cohort of 69 patients. For CRC liver metastasis, Morpheus discovered two cohort-dependent therapeutic strategies consisting of blocking different subsets of CXCR4, PD-1, PD-L1 and CYR61 that are predicted to improve T-cell infiltration in a cohort of 30 patients. Our work provides a paradigm for counterfactual-based prediction and design of cancer therapeutics based on classification of immune system activity in spatial omics data.

Results

Counterfactual optimization for therapeutic prediction

The general logic of Morpheus (Figure 1A) is to first train, in a self-supervised manner, a classifier to predict the presence of CD8+ T cells from multiplexed tissue images (Figure 1B). Then we compute counterfactual instances of the data by performing gradient descent on the input image, allowing us to discover perturbations to the tumor image that increases the classifier's predicted likelihood of CD8+ T cells being present (Figure 1C). The altered image represents a perturbation of the TME predicted to improve T-cell infiltration. We mask CD8+ T cells from all images to prevent the classifier from simply memorizing T-cell expression patterns, guiding it instead to learn environmental features indicative of T-cell presence.

We leverage IMC profiles of human tumors to train a classifier to predict the spatial distribution of CD8+ T cell in a self-supervised manner. Consider a set of images $\{I^{(i)}\}$, obtained by dividing IMC profiles of tumor sections into local patches of tissue signaling environments, where $I^{(i)} \in \mathbb{R}^{l \times w \times c}$ is an array with l and w denoting the pixel length and width of the image and c denoting the number of molecular channels in the images (Figure 1B). Each image shows the level of c proteins across all cells within a small patch of tissue. From patch $I^{(i)}$, we obtain a binary label $s^{(i)}$ indicating the presence and absence of CD8+ T cells in the patch and a masked copy $x^{(i)}$ with all signals originating from CD8+ T cells removed (see Methods). The task for the model f is to classify whether T cells are present ($s^{(i)} = 1$) or absent ($s^{(i)} = 0$) in image $I^{(i)}$ using only its masked copy $x^{(i)}$. Specifically, $f(x^{(i)}) \in [0, 1]$ is the predicted probability of T cells, and then we apply a classification threshold p to convert this probability to a predicted label $\hat{s}^{(i)} \in \{0, 1\}$. Since we obtain the image label $s^{(i)}$ from the image $I^{(i)}$ itself by unsupervised clustering of individual cells, our overall task is inherently self-supervised.

Given a set of image patches, we train a model f to minimize the following T cell prediction loss, also known as the binary cross entropy (BCE) loss,

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[s^{(i)}\log\left(\hat{s}^{(i)}\right) + \left(1 - s^{(i)}\right)\log\left(1 - \hat{s}^{(i)}\right)\right], \tag{1}$$

where

$$\hat{s}^{(i)} = \begin{cases} 1 & \text{if } f(x^{(i)}) \geq p \\ 0 & \text{if } f(x^{(i)}) < p \end{cases} \tag{2}$$

and p is the classification threshold. We select p by minimizing the following root mean squared error (RMSE) on a separate set of tissue sections Ω,

$$\text{RMSE}^2 = \frac{1}{|\Omega|}\sum_{j \in \Omega}\left|\frac{1}{N_j}\sum_{i=1}^{N_j} s^{(i)} - \frac{1}{N_j}\sum_{i=1}^{N_j}\hat{s}^{(i)}\right|^2. \tag{3}$$

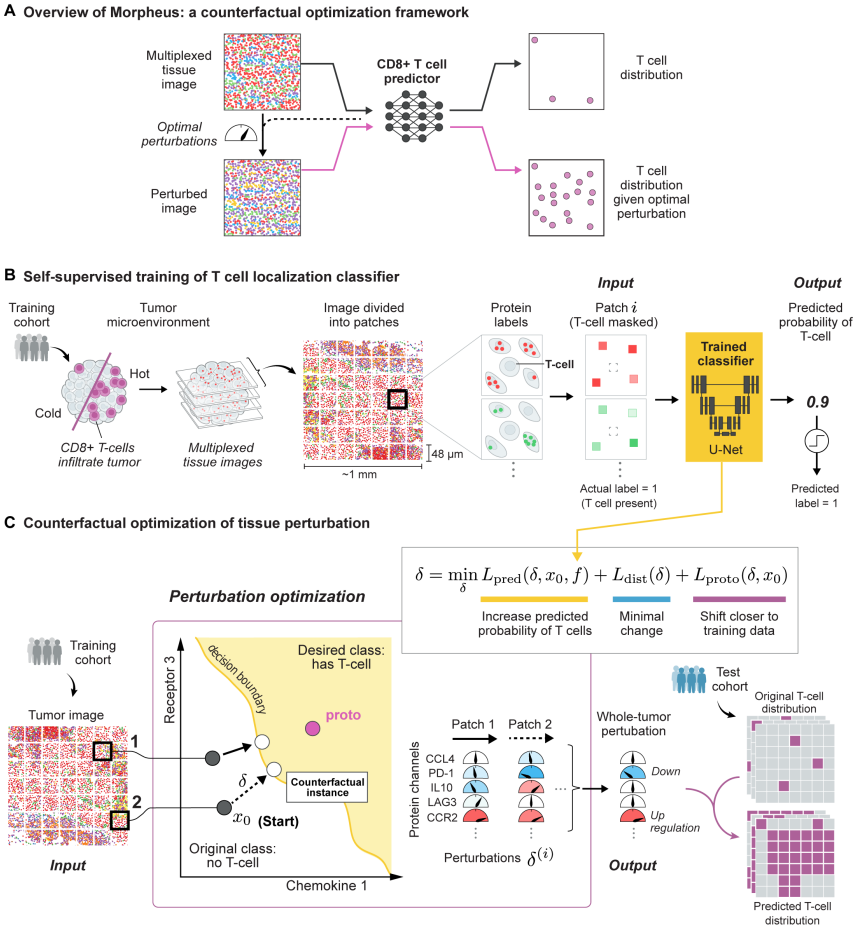


Fig. 1: **An integrated counterfactual optimization framework for discovering therapeutic strategies predicted to drive CD8+ T cell infiltration in human tumors.** (A) Overview of the Morpheus framework, which consists of first (B) training a neural network classifier to predict the presence of CD8+ T cells from multiplexed tissue images where CD8+ T cells are masked. (C) The trained classifier is then used to compute an optimal perturbation vector $\delta^{(i)}$ per patch by jointly minimizing three loss terms (L_{pred}, L_{dist}, L_{proto}). The perturbation $\delta^{(i)}$ represents a strategy for altering the level of a small number of signaling molecules in patch $x_0^{(i)}$ in a way that increases the probability of T cell presence as predicted by the classifier. The optimization also favors perturbations that shift the image patch to be more similar to its nearest T-cell patches in the training data, shown as proto. Each perturbation corresponds to adjusting the relative intensity of each imaging channel. Taking the median across all perturbations produces a whole-tumor perturbation strategy, which we assess by perturbing *in silico* tumor images from a test patient cohort and examining the predicted T cell distribution after perturbation.

The RMSE is a measure of the differences between the observed and predicted proportions of T cell patches in a tissue section averaged across a set of tissues Ω, which we take to be the validation set.

We evaluated the performance of various classifiers, including both traditional convolutional neural networks (CNNs) and vision transformers. In all cases, we observed similar performance (Table S3). We settled on a U-Net architecture because of ease of extension of the model to multichannel data sets. Our U-Net classifier consists of a standard U-Net architecture [29] and a fully-connected layer with softmax activation (Methods). To increase the number of samples available for training, we take advantage of the spatial heterogeneity of TMEs and divide each tissue image into 48 µm × 48 µm patches upon which the classifier is trained to predict T cell presence (Methods).

Using our trained classifier and IMC images of tumors, we employ a counterfactual optimization method to predict tumor perturbations that enhance CD8+ T cell infiltration (Figure 1C). For each image patch $x_0^{(i)}$ lacking CD8+ T cell, our optimization algorithm searches for a perturbation $\delta^{(i)}$ such that our classifier f predicts the perturbed patch $x_p^{(i)} = x_0^{(i)} + \delta^{(i)}$ as having T cells, hence $x_p^{(i)}$ is referred to as a counterfactual instance. Ideally, we also want our perturbation to be minimal in that it only requires targeting a small number of molecule, and realistic in that the counterfactual instance is not far from image patches in our training data so we can be more confident of the model's prediction. We can obtain a perturbation $\delta^{(i)}$ with these desired properties by solving the following optimization problem adopted from [30],

$$\delta^{(i)} = \min_{\delta} L_{\text{pred}}(x_0^{(i)}, \delta) + L_{\text{dist}}(\delta) + L_{\text{proto}}(x_0^{(i)}, \delta), \tag{4}$$

such that

$$\begin{aligned} L_{\text{pred}}(x_0^{(i)}, \delta) &= c\max(-f(x_0^{(i)} + \delta),\, -p), \\ L_{\text{dist}}(\delta) &= \beta\|\delta\|_1 + \|\delta\|_2^2, \\ L_{\text{proto}}(x_0^{(i)}, \delta) &= \theta\|x_0^{(i)} + \delta - \text{proto}^{(i)}\|_2^2 \end{aligned} \tag{5}$$

where $\delta^{(i)}$ is a 3D tensor that describes perturbation made to each pixel of the patch.

The three loss terms in Equation (4) each correspond to a desirable property of the perturbation we aim to discover. The term L_{pred} encourages validity, in that the perturbation increases the classifier's predicted probability of T cells to be larger than p, so the network will predict the perturbed tissue patch as having T cells when it previously did not contain T cells. Next, the term L_{dist} encourages sparsity, in that the perturbation does not require making many changes to the TME, by minimizing the distance between the original patch $x_0^{(i)}$ and the perturbed patch $x_{\text{p}}^{(i)} = x_0^{(i)} + \delta$ using elastic net regularization. Lastly, the term $\text{proto}^{(i)}$ in the expression for L_{proto} refers to the nearest neighbour of $x_0^{(i)}$ among all patches in the training set that are classified as

having T cells (see Methods). Thus the term L_{proto} explicitly guides the perturbed image $x_{\text{p}}^{(i)}$ to lie close to the data manifold defined by our training set, making perturbed patches appear similar to what has been observed in TMEs infiltrated by T cells.

Since drug treatments cannot act at the spatial resolution of individual micron-scale pixels, we constrain our search space to only perturbations that affect all cells in the image uniformly. Specifically, we only search for perturbations that change the level of any molecule by the same relative amount across all cells in an image. We incorporate this constraint by defining $\delta^{(i)}$ in the following way,

$$\delta^{(i)} = \gamma^{(i)} \odot_3 x_0^{(i)}, \tag{6}$$

where $\gamma^{(i)} \in \mathbb{R}^c$ defines a single factor for each channel in the image and the circled dot operator represent channel-wise multiplication, so that within each channel, the scaling factor is constant across the spatial dimensions of the image. In practice, we directly optimize for $\gamma^{(i)}$, where $\gamma_j^{(i)}$ can be interpreted as the relative change to the mean intensity of the j-th channel. However, given our classifier does have fine spatial resolution, we can search for targeted therapies such as perturbing only a specific cell type or restricting the perturbation to specific tissue locations by changing Equation (6) to match these different types of perturbation.

Taken together, our algorithm obtains an altered image predicted to contain T cells from an original image which lacks T cells, by minimally perturbing the original image in the direction of the nearest training patch containing T cells until the classifier predicts the perturbed image to contain T cells. Since our strategy may find different perturbations for different tumor patches, we reduce the set of patch-wise perturbations $\{\delta^{(i)}\}_i$ to a whole-tumor perturbation by taking the median across the entire set.

Convolutional neural networks predict T-cell distribution

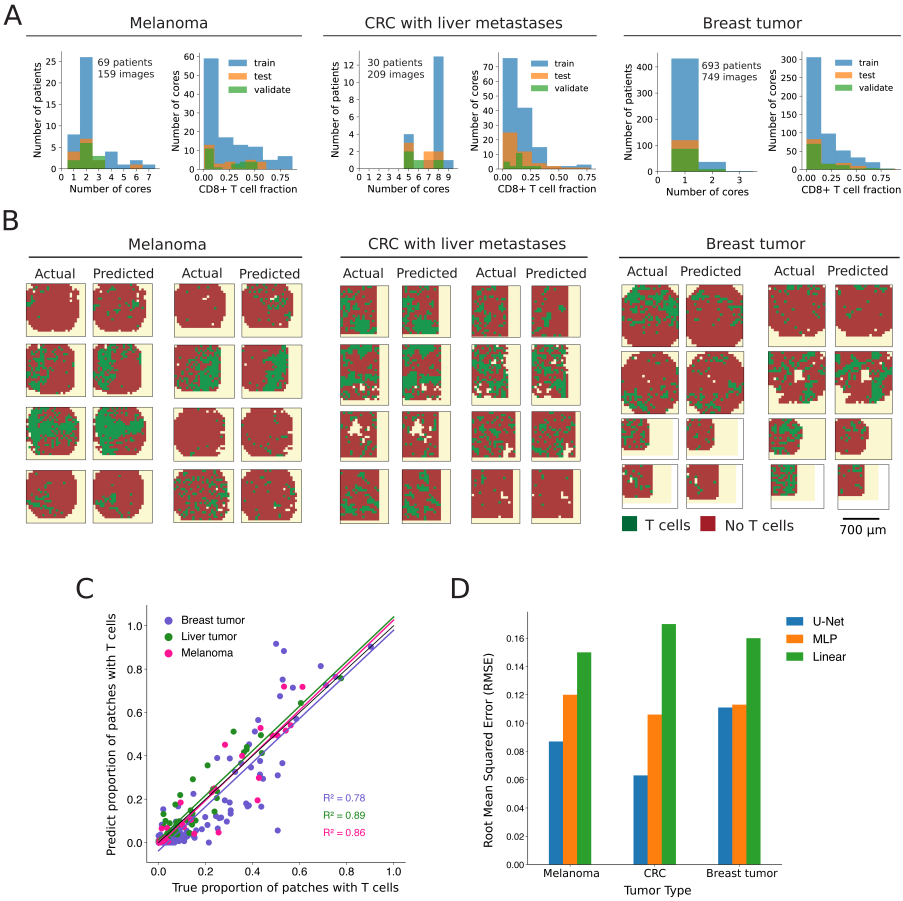


Fig. 2: **U-Net classifiers accurately predict T cell distribution in IMC images of melanoma, metastatic liver, and breast tumor.** (A) Histograms showing the distribution of tumor cores per patient and CD8+ T cell fractions per core across all three data sets and data splits. (B) Predicted and actual T cell distribution of tissue sections from test cohorts in melanoma, liver tumor, and breast tumor data set. (C) Predicted and true proportion of patches with T cells within a tissue section, each dot corresponds to a tissue section, diagonal black line indicates perfect prediction. (D) The RMSE (Equation (3)) across all (test) tissue sections for three different classes of models.

We applied Morpheus to two publicly available IMC data sets of tumors from patients with metastatic melanoma [26] and colorectal cancer (CRC) with liver metastases [27] (Figure 2A). We validate the infiltration prediction on an

additional breast cancer data set [28]. While this breast cancer data focuses on cell type markers over functional modulators of T-cell infiltration, making it unsuitable for therapeutic prediction, it serves to further validate our ML-based prediction of T-cell infiltration.

The melanoma data set [26] was obtained by IMC imaging of 159 tumor cores from 69 patients with stage III or IV metastatic melanoma. Each tissue was imaged across 39 molecular channels, consisting of markers for tumor, immune, and stromal cells, as well as 11 different chemokines (RNA) (Methods). The CRC data set [27] consists of 209 tissue sections taken from 30 patients imaged across 42 channels, including 60 sections from primary CRC tumors, 89 sections CRC metastases to the liver and 60 "healthy" liver sections obtained away from the metastases (Methods). The breast cancer data set [28] was obtained by IMC imaging of 749 breast tumor cores from 693 patients. The tissues were imaged across 37 channels, consisting of markers for tumor, lymphoid, myeloid and stromal cells (Methods).

For each of the three tumor data sets, we trained a separate U-Net classifier that effectively predicts CD8+ T cell infiltration level in unseen tumor sections (Methods). The two classifiers trained on melanoma and CRC data sets achieved the best performance with an AUROC of 0.77 and 0.8 respectively, whereas the classifier trained on breast tumors achieved a AUROC of 0.71 (Table S2). Figure 2B shows examples of actual and predicted T cell distributions in tumor sections. For each tissue section of a cancer type, the predictions were obtained by applying the corresponding U-Net classifier to each image patch independently. By visual inspection, our classifiers consistently captures the general distribution of T cells. Comparing the true proportion of T-cell patches in a tissue section against our predicted proportion also shows strong agreement (Figure 2C). The true proportion of patches with T cells is calculated by dividing the number of patches within a tissue section that contain CD8+ T cells by the total number of patches within that section. We quantify the performance of our U-Nets on the entire test data set using the RMSE (Equation (3)), which represents the mean difference between our predicted proportion and the true proportion per tumor section (Figure 2D). Our classifiers performs well on liver tumor and melanoma, achieving a RMSE of only 6% and 8% respectively and a relatively lower performance of 11% on breast tumor. Taken together, these results suggest that our classifier can accurately predict the T cell infiltration status of multiple tumor types.

In order to gain insight into the relative importance of non-linearity and spatial information in the performance of the U-Net on the T cell clasification task, we compared the U-nets' performance to a logistic regression model (LR) and a multi-layer perceptron (MLP). Both the LR and MLP model are given only mean channel intensities as input, so neither have explicit spatial information. Furthermore, the LR model is a linear model with a threshold whereas the MLP is a non-linear model. Figure 2D shows that across all three cancer data sets, the MLP classifier consistently outperforms the logistic regression model, reducing RMSE by $20-40\%$ to suggest that there are significant non-linear interactions between different molecular features in terms of their effect

on T cell localization. The importance of spatial features on the T cell prediction task, however, is less consistent across cancer types. Figure 2D shows that for predicting T cells in breast tumor, the U-Net model offers negligible boost in performance relative to the MLP model ($<$ 2% RMSE reduction), whereas for liver tumor, the U-Net model achieved a RMSE 50% lower compared to the MLP model. This result suggests that the spatial organization of signals may have a stronger influence on CD8+ T cell localization in liver tumor compared to breast tumor.

246 # Applying Morpheus to metastatic melanoma samples

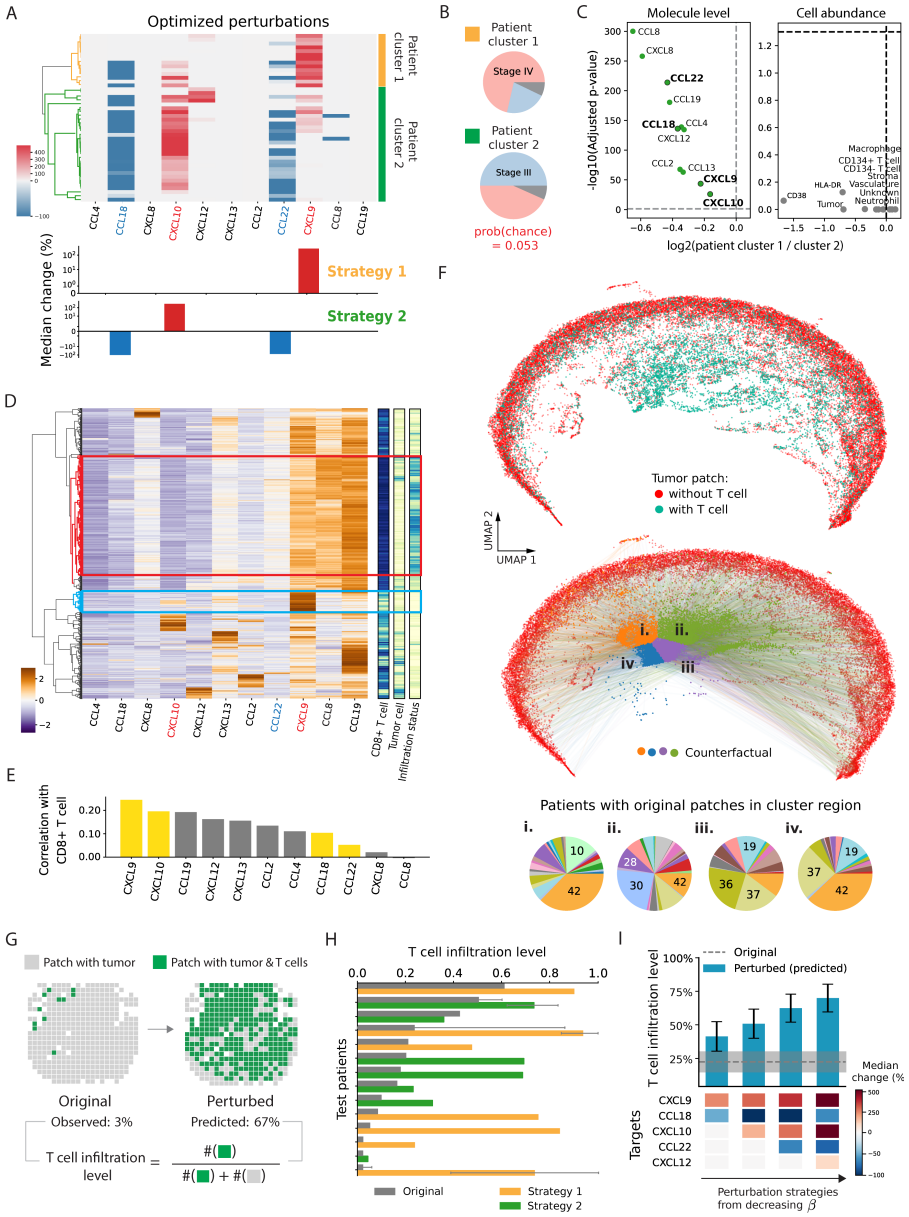


Fig. 3: **Combinatorial chemokine therapy predicted to drive T cell infiltration in patients with metastatic melanoma** (A) Whole-tumor perturbations optimized across IMC images of patients (row) from the training cohort, with bar graph showing the median relative change in intensity for each molecule.

Fig. 3: **(continued)** (B) Distribution of cancer stages among patients within two clusters, gray indicates unknown stage, chance probability from hypergeometric distribution. (C) Volcano plot comparing chemokine level and cell type abundance from patient cluster 1 and 2, computed using mean values and Wilcoxon rank sum test. Gray indicates non-statistical significance. (D) Patch-wise chemokine profile (left); 1-D heatmap (right): infiltration status (light/dark = from infiltrated/deserted tumor), tumor cell (light/dark = present/absent), CD8+ T cells (light/dark = present/absent). (E) Patch-wise correlation between chemokine signals and the presence of CD8+ T cells. (F) (Top) UMAP projection of tumor patches (chemokine channels) show a clear separation of masked patches with and without T cells. (Bottom) colored arrows connect UMAP projection of patches without T cells and their corresponding counterfactual (perturbed) patch, where the colors correspond to k-nearest neighbor clusters (i-iv) of the counterfactual patches, highlighting the minimal nature of the perturbations. Pie charts (i-iv) shows the distribution of patients whose original tumor patches are found in the corresponding cluster regions in the UMAP. (G) Cell maps computed from a patient's IMC image, showing the distribution of T cells before and after perturbation. (H) Original vs. perturbed (predicted) mean infiltration level across all patients (test cohort) with 95% confidence interval (only shown for patients with more than 2 samples). Stage IV patients received perturbation strategy 1 (yellow), stage III patients received perturbation strategy 2 (green). (I) Mean infiltration level across all patients (test cohort) for optimized perturbation strategies of varying sparsity, error bar represents 95% CI.

Applying our counterfactual optimization procedure using the U-Net classifier trained on melanoma IMC images, we discovered a combinatorial therapy predicted to be highly effective in improving T cell infiltration in patients with melanoma. We restricted the optimization algorithm to only perturb the level of chemokines, which are a family of secreted proteins that are known for their ability to stimulate cell migration [31] and have already been harnessed to augment T-cell therapy [32]. By optimizing over multiple chemokines, Morpheus opens the door to combinatorial chemokine therapeutics that has the potentially to more effectively enhance T cell infiltration into tumors. Figure 3A shows that patients from the training cohort separate into two clusters based on hierarchical clustering of perturbations computed for each patient. Taking median across all patients in cluster 1, the optimized perturbation is to increase CXCL9 level by 370%, whereas in patient cluster 2, the optimized perturbation consists of increasing CXCL10 level by 280% while decreasing CCL18 and CCL22 levels by 100% and 70% respectively (Figure 3A). Both CXCL9 and CXCL10 are well-known for playing a role in the recruitment of CD8+ T cells to tumors. On the other hand, CCL22 is known to be a key chemokine for recruiting regulatory T cells [33] and CCL18 is known to induce an M2-macrophage phenotype [34], so their expression likely promotes

an immunosuppressive microenvironment inhibitory to T cell infiltration and function.

Figure 3B shows that the choice of which of these two strategies was selected for a patient appears to be strongly associated with the patient's cancer stage, with strategy 1 being significantly enriched for patients with stage IV metastatic melanoma and strategy 2 being significantly enriched for patients with stage III cancer, with a probability of 0.053 of such difference being due to chance. Probing deeper into the difference between these two patient clusters, we find that all chemokines have lower mean expression in the tumors of patients in cluster 1 compared to cluster 2, while there are no significant differences between the two groups in terms of the cell type compositions within tumors (Figure 3C). Since the levels of CCL22 and CCL18 is 37% and 31% higher in patients from cluster 2 and both chemokines have been implicated in having an inhibitory effect on T-cell infiltration, it is reasonable that the optimization algorithm suggests inhibiting CCL18 and CCL22 only for patients in cluster 2. However, the switch from boosting CXCL9 to CXCL10 is not as straightforward. A possible explanations is that boosting CXCL10 is important when blocking CCL18 and CCL22 in order for the perturbed patches to stay close to the data manifold, leading to more realistic tissue environments.

Morpheus selected perturbations that would make the chemokine composition of a TME more similar to T cell rich regions of immune-infiltrated tumors. Figure 3D shows that melanoma tissue patches can be clustered into distinct groups based on their chemokine concentration profile. One cluster (highlighted in blue) contains exactly the patches from immune-infiltrated tumors that contain both tumor and T cells, which likely represents a chemokine signature that is suitable for T cell infiltration. Alternately, a second cluster (highlighted in red) which contains patches from immune-desert tumors that have tumor cells but no T cells likely represents an unfavorable chemokine signature. In comparison to the cluster highlighted in red, Figure 3D shows the cluster highlighted in blue contains elevated levels of CXCL9, CXCL10 and reduced levels of CCL22 which partially agrees with the perturbation strategy (Figure 3A) discovered by Morpheus. Lastly, Figure 3E shows that our four selected chemokine targets cannot simply be predicted from correlation of chemokine levels with the presence of CD8+ T cells, as both CCL18 and CCL22 are weakly correlated (< 0.1) with CD8+ T cells even though the optimized perturbations requires inhibiting both chemokines, suggesting the presence of significant nonlinear effects not captured by correlations alone.

We can directly observe how Morpheus searches for efficient perturbations by viewing both the original patch and perturbed patches in a dimensionally-reduced space. Figure 3F (top) shows a UMAP projection where each point represents the chemokine profile of an IMC patch. T-cell patches (with their CD8+ T cells masked) are well-separated from patches without CD8+ T cells. The colored arrows in the bottom UMAP of Figure 3F illustrate the perturbation for each patch as computed by Morpheus, and demonstrate two key

features of our algorithm. First, optimized perturbations push patches without T cells towards the region in UMAP space occupied by T-cell-infiltrated patches. Second, the arrows in Figure 3C are colored to show that optimized perturbations seem efficient in that patches are perturbed just far enough to land in the desired region of space. Specifically, red points that start out on the right edge end up closer to the right after perturbation (region ii and iii), while points that start on the left/bottom edge end up closer to the left/bottom (region i and iv), respectively. We make this observation while noting that UMAP, though designed to preserve the topological structure of the data, is not a strictly distance-preserving transformation [35]. Furthermore, the pie charts (i-iv) are colored by the patient of origin to show the region of space where points are being perturbed to are not occupied by tissue samples from a single patient with highly infiltrated tumor. Rather, these regions consist of tissue samples from multiple patients, suggesting that our optimization procedure can synthesize information from different patients when searching for therapeutic strategies.

After applying the second perturbation strategy from Figure 3A *in silico* to IMC images of a tumor, Figure 3G shows that T cell infiltration level (defined as the proportion of tumor patches with T cells) is predicted to increase by 20 fold. We applied our two perturbation strategies on patients in our test cohort *in silico* after stratifying by cancer stage, using strategy 1 on patients with stage IV melanoma and strategy 2 on patients with stage III melanoma.Figure 3H shows that this predicted improvement holds across nearly all 14 patients from the test group, boosting T cell infiltration level from an average of 23% across samples to a predicted 63% post perturbation. For the three test patients with multiple tumor sections (patient 64, 57, 89), we see small to moderate variation in predicted improvement across samples.

The combinatorial nature of our optimized perturbation strategy is crucial to its predicted effectiveness. We systematically explored the importance of combinatorial perturbation by changing parameter β of Equation (4) which adjusts the sparsity of the strategy, where a more sparse strategy means fewer molecules are perturbed. Figure 3I shows that perturbing multiple targets is predicted to be necessary for driving significant T cell infiltration across multiple patients, with the best perturbation strategy involving two targets predicted to generate only 60% of the infiltration level achieved by the best perturbation strategy involving four targets. In conclusion, within the scope of the chemokine targets considered, combinatorial perturbation of the TME appears necessary for improving T cell infiltration in metastatic melanoma.

Applying Morpheus to CRC with liver metastases samples

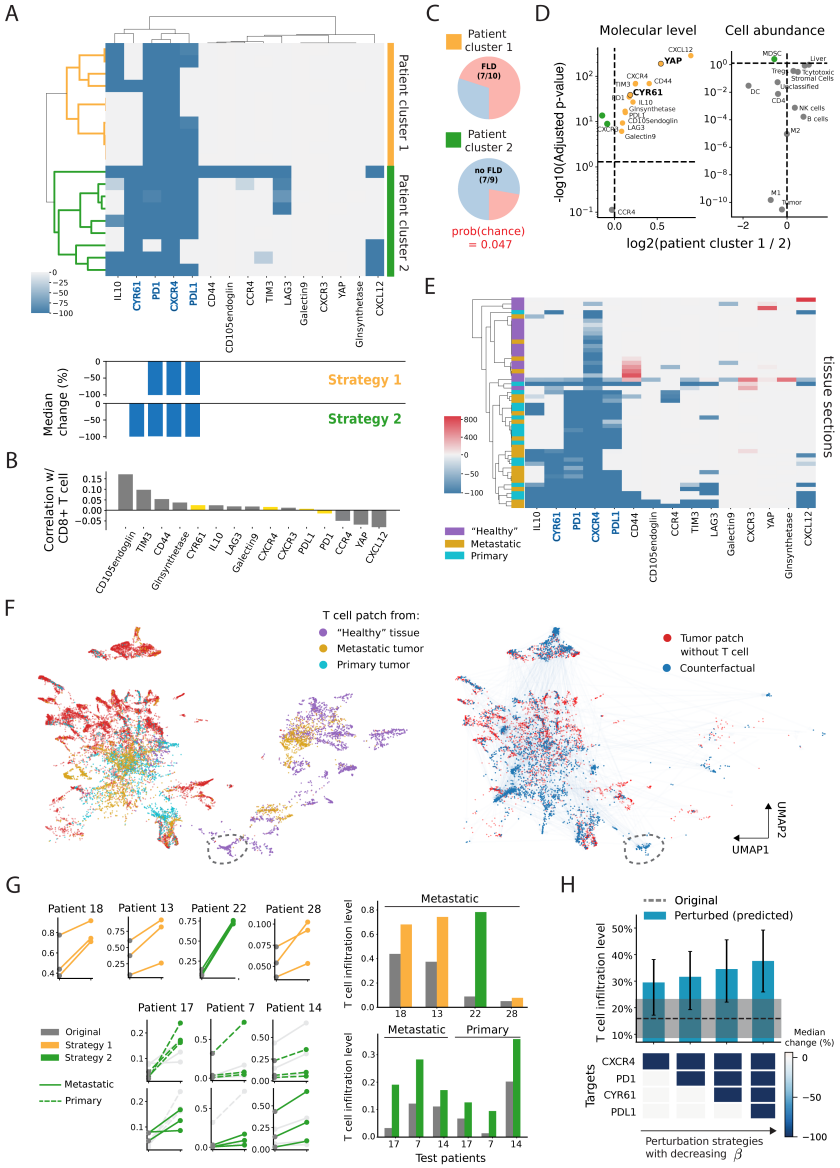


Fig. 4: **Blocking subsets of PD-L1, CXCR4, PD-1, and CYR61 predicted to drive T cell infiltration in CRC cohort.** (A) Optimized tumor perturbations aggregated to the patient (row) level (train cohort). Bar graph shows the median relative change in intensity for each molecule across all patients within their cluster.

Fig. 4: **(continued)** (B) Patch-wise correlation between the levels of different molecules and the presence of CD8+ T cells. (C) Pie charts show proportion of patients in each cluster that have fatty liver disease (FLD), chance probability from hypergeometric distribution. (D) Volcano plot comparing molecule levels and cell type abundance between the two patient cluster using tumor tissues, computed using mean values and Wilcoxon rank sum test with Bonferroni correction. (E) Optimized perturbations aggregated to the level of tissue samples (row). (F) UMAP projection of IMC patches, left UMAP shows T cell patches colored by the tissue samples they are taken from. right UMAP shows counterfactual (perturbed) instances optimized for tumor patches without T cells (red). (G) Line plots shows T-cell infiltration level for each tissue section from the test cohort, before and after perturbation. Bar plots show predicted mean T-cell infiltration level for each test patient. (H) Mean infiltration level across all test patients using perturbation strategies of varying sparsity, obtained by varying β in Equation (4), error bar represents 95% CI.

Applying Morpheus to IMC images from the CRC cohort, we discovered two patient-dependent therapies predicted to be highly effective in improving T cell infiltration. Figure 4A shows the optimal perturbations computed for every patient from the training cohort, aggregated over all tumor samples for each patient. Our method consistently discovered two distinct patient-dependent strategies for improving T cell infiltration, as revealed by hierarchical clustering of all patient-level perturbations (Figure 4A). Taking median over patients in the first cluster, the optimized strategy involves completely inhibiting PD-1, PD-L1, and CXCR4. While for the second group of patients, the optimized strategy involves completely inhibiting CYR61, PD-1, PD-L1, and CXCR4 (Figure 4A). Interestingly, all four of the perturbation targets correlated poorly with the presence of CD8+ T cells compared to the other proteins that were not selected as perturbation targets (Figure 4B), suggesting the presence of significant spatial and nonlinear effects not captured by correlations alone.

All perturbation targets identified by our optimization procedure have been found to play crucial roles in suppressing T cell function in the TME, and treating patients with inhibitors against subsets of the selected targets have already improved T cell infiltration in human CRC liver metastases. Regulatory T cells (Tregs) are recruited into tumor through CXCL12/CXCR4 interaction [36], and the PD-1/PD-L1 pathway inhibits CD8+ T cell activity and infiltration in tumors. In addition, CYR61 is a chemoattractant and was recently shown to drive M2 TAM infiltration in patients with CRC liver metastases [27]. Inhibition of both PD-1 and CXCR4, which were consistently selected by our algorithm as targets, have already been shown to increase CD8+ T cell infiltration in both patients with CRC and mouse models [37–39]. Finally, Figure 4A shows that the fifth most common proposed perturbation involves inhibiting IL-10. Indeed, blockade of IL-10 was recently shown to increase the

frequency of non-exhausted CD8+ T cell infiltration in slice cultures of human CRC liver metastases [40].

The emergence of the two distinct perturbation strategies may be explained by variation in liver fat build-up among patients. Patient cluster 1 is made up of significantly more patients with fatty liver disease (70% FLD) compared to patient cluster 2 (22%), where the probability of this due purely to chance is 0.047 (Figure 4C). Furthermore, Figure 4D shows that both YAP and CYR61 levels are significantly higher in tumors from patient cluster 1, by 50% and 15% respectively. Indeed, CYR61 is known to be associated with non-alcoholic fatty liver disease [27] and YAP is a transcription coregulator that induces CYR61 expression [41]. However despite patients in cluster 1 having higher levels of CYR61, it is only for patients in cluster 2 where the optimal strategy involves blocking CYR61. We postulate that this seemingly paradoxical finding may arise because removing CYR61 from patients in cluster 1 represents a more pronounced perturbation, given their inherently higher concentration. A perturbation of this magnitude would likely shift the tumor profile significantly away from the data manifold, where the classifier's prediction about the perturbation's effect becomes less reliable, hence such a perturbation would be heavily penalized during optimization due to the L_{proto} term.

Using only raw image patches, Morpheus discovers tissue-dependent perturbation strategies (Figure 4E). As depicted in Figure 4E, by aggregating perturbations at the individual tissue level, we observe that the optimized perturbation for "healthy" liver sections is straightforward, necessitating only the inhibition of CXCR4. Recall "healthy" sections are samples obtained away from sites of metastasis. In contrast, promoting T cell infiltration into primary colon tumors is anticipated to involve targeting a minimum of three signals. Our method finds that liver metastases appears to fall between these two tissue types. The optimized perturbation strategy for some liver metastases samples is to block CXCR4, while requiring the inhibition of the same set of signals as primary tumors for others. Furthermore, direct comparison between perturbations optimized for metastatic tumor and primary tumor samples does not reveal a significant difference in strategy (Figure S2). We can partly understand the discrepancy between tissues by plotting a UMAP projection of all T cell patches from the three tissue types (Figure 4F, left). The clear separation between T cell patches from "healthy" tissue and those from primary tumors underscores that the signaling compositions driving T cell infiltration likely differ substantially between the two tissue types. This distinction is likely what prompted our method to identify markedly different perturbation strategies. Furthermore, some patches from metastatic tumors co-localize with "healthy" tissue patches in UMAP space, while other patches co-localizes with primary tumor patches. This observation again aligns with our previous result, where optimized perturbations for metastases samples can bear similarities to strategies for either "healthy" tissue or primary tumor (Figure 4E).

Despite the CRC data set comprising a complex blend of healthy, tumor, and hybrid metastatic samples, Morpheus targets the most pertinent tissue

type when optimizing perturbations. During both the model training and counterfactual optimization phases, we did not make specific efforts to segregate the three tissue types. Furthermore, we did not provide tissue type labels or any metadata. Despite these nuances, Figure 4F shows that the counterfactual instances for tumor patches (dark blue) from primary and metastases samples are mostly perturbed to be near T cell patches from primary (cyan) and metastatic tumor (gold), instead of being perturbed to be similar to T cell patches from "healthy" tumors (purple). This result is partly a consequence of our prototypical constraint which encourages patches to be perturbed towards the closest T-cell patch. For a patch from a metastatic tumor without T cells, the closest (most similar) T cell patch is likely also from a metastatic tumor than from a "healthy" tissue. However, there are occasional exceptions where T cell patches from "healthy" tissues can influence the optimization of tumor tissues, as outlined by the dashed ellipse in Figure 4F, especially if they share similar features as tumor regions.

The two therapeutic strategies we discovered generalize to patients in our test cohort (Figure 4G,H). Given that we have two therapeutic strategies, one enriched for patients with FLD and another for patients without FLD, we apply different perturbation strategies *in silico* across all test patients depending on their FLD status. Aggregated to the patient level, Figure 4G shows that CD8+ T cell infiltration level is predicted to increase for nearly all patients, with the exception of patient 28. Furthermore, aggregating to the entire test cohort, Figure 4H shows a statistically significant boost to mean infiltration level from 15% to a predicted 35% post perturbation. However, when comparing individual tissue samples, Figure 4G reveals significant variation in the predicted response to perturbation among samples from the same patient and tissue types. In patient 7, one primary tumor sample is predicted to see a nearly three-fold increase in T cell infiltration after perturbation, yet almost no change is expected for patient 7's other two primary and three metastatic samples. Similar patterns are observed in patients 14 and 17. This marked variability in response among a significant portion of test patients underscores the challenges posed by intra-tumor and inter-patient heterogeneity in devising therapies for CRC with liver metastases. This result further implies that, for studying CRC with liver metastases, collecting numerous tumor sections per patient could be as crucial as establishing a large patient cohort. Lastly, combinatorial perturbation is again predicted to be necessary to drive significant T-cell infiltration in large patient cohorts. By increasing β in Equation (4), we generated strategies with between one and four total targets, where our four-target perturbation is the only strategy predicted to produce a statistically significant boost to T-cell infiltration (Figure 4H).

Discussion

Our integrated deep learning framework, Morpheus, combines deep learning with counterfactual optimization to directly predict therapeutic strategies

from spatial omics data. One of the major strengths of Morpheus is that it scales efficiently to deal with large diverse sets of patients samples including metachronous tissue from the same patients but different sites, which will be crucial as more spatial transcriptomics and proteomics data sets are quickly becoming available [42]. Larger data sets could allow us to train more complex models such as vision transformers, capturing long range interactions in tissues to improve prediction of T-cell localization. Furthermore, a large set of diverse patient samples will more accurately capture the extent of tumor heterogeneity, enabling Morpheus to discover therapeutic strategies for different sub-classes of patients.

For future work, we would like to apply Morpheus to spatial transcriptomics data sets with hundreds to thousands of molecular channels. Although spatial transcriptomics can profile significantly more molecules compared to spatial proteomic techniques [15, 16], the number of spatial transcriptomic profiles of human tumors is currently limited due to the cost, with most public data sets containing single tissue sections from 1-5 patients which is far too small to apply Morpheus. However, spatial transcriptomics is likely to be more standardized compared to proteomics, which use customized panels. As commercial platforms for spatial transcriptomics start to come online [43], we will likely be seeing large scale spatial transcriptomics data sets in the near future, with $\sim$ 70-90% of the same probes shared between experiments.

A technical extension of Morpheus involves incorporating prior knowledge of gene-gene interactions to model the causal relations between genes. Molecular features in tissue profiles can exhibit strong dependencies, therefore, changing the level of one molecule can affect the expression of others. For example, increased levels of interferon-gamma (IFN-γ) in the tumor microenvironment, can upregulate the expression of PD-L1 on tumor cells [44]. In order to be more realistic and actionable, a counterfactual should maintain these known causal relations. We can apply a regularizer to penalize counterfactuals that are less feasible according to established gene interactions from knowledge graphs, such as Gene Ontology [45].

Other extensions of Morpheus includes predicting cell-type specific perturbations, which can be done by directly restricting the perturbation to only alter signals within specific cell types. Additionally, although we applied Morpheus to the specific problem of driving T cells to infiltrate solid tumors, we can generalize our framework to predict candidate therapeutics that alter the localization of other cell types. For example, Morpheus can train a classifier model to predict localization of TAMs and compute perturbations predicted to reduce their abundance in the TME.

In this work, we focused on identifying generalized therapies by pooling predictions across multiple patient samples, but we can also apply Morpheus to find personalized therapy for treating individual patients. The variation in the optimized perturbations we observe among patients in both melanoma and liver data sets suggest personalize treatments could be significantly more effective compared to generalized therapies (Figure 3A, Figure 4A). Furthermore,

Figure 4G shows that a therapeutic strategy could have highly variable effect even across different tissue samples from the same patient. This variability suggests that to generate therapy for an individual patient, it may be necessary to acquire significant quantities of biopsy data. We can then apply our optimization procedure to a random subset of the samples, and then test the resulting perturbation strategy on the remaining samples to see how well the strategy is predicted to perform across an entire tumor or other primary/secondary tumors.

Incorporating Morpheus in a closed loop with experimental data collection is another promising direction for future work. Data can be collected from patients or animal models with perturbed/engineered signaling context, and this data can be easily fed back into the classifier model to refine the model's prediction. The perturbation could be based on what the model predicts to be effective interventions, as is the case with Morpheus. We can also study tissue samples on which the model tends to make the most mistake and train the model specifically using samples from similar sources, such as similar patient strata or disease state.

Methods

IMC data sets

All data sets used in this paper are publicly available. Metastatic melanoma data set from Hoch et al. [26] contains 159 images or cores taken from 69 patients, collected from sites including skin and lymph-node. CRC liver metastases data set from Wang et al. [27] contains 209 images or cores taken from 30 patients. Breast tumor data set from Danenberg et al. [28] contains 693 images or cores taken from 693 patients. The RNA and protein panels used for each of the three data sets are listed in Table 1.

Data split

For all three IMC data sets, we followed the same data splitting scheme to divide patients into three different groups (training, validation, testing) while ensuring similar class balance across the groups, which in our case means that the proportion of image patches with and without T cells are roughly equal across the three groups for each data set. Specifically, each image within a data set was divided into 48 µm × 48 µm patches and the number of patches with and without CD8+ T cells was computed for each image. Furthermore, each patch was downsampled from 48 × 48 pixels to 16 × 16 pixel dimension where each pixel now represents a 3 µm × 3 µm region. We applied spectral analysis to study the effect of using different patch size to predict T cell infiltration and found that our selected patch size remains highly informative of T cell presence (Figure S1). Next, the patients are shuffled between the three groups until three criteria are met: 1) the number of patients across the three groups follow a 65/15/20 ratio, 2) the difference in class proportion between any two

Metastatic melanoma		CRC with liver metastases		Breast tumor	
Vimentin	**DapB**	CD45	Glnsynthetase	Histone H3	SMA
CD163	**CCL4**	CD163	NKG2D	CK5	CD38
B2M	**CCL18**	CCR4	PD-L1	HLA-DR	CK8-18
CD134	**CXCL8**	FAP	CD11c	CD15	FSP1
CD68	**CXCL10**	LAG3	HepPar1	CD163	ICOS
GLUT1	**CXCL12**	FOXP3	αSMA	OX40	CD68
CD3	**CXCL13**	CD4	CD105	HER2 (3B5)	CD3
LAG3	**CCL2**	CD68	VISTA	Podoplanin	CD11c
PD-1	**CCL22**	CD20	CD8α	PD-1	GITR
HistoneH3	**CXCL9**	TIM3	CXCR4	CD16	c-Caspase3
CCR2	**CCL19**	PD-1	iNOS	CD45RA	B2M
PD-L1	**CCL8**	CD31	CYR61	CD45RO	FOXP3
CD8	SMA	CDX2	CAIX	CD20	ER
SOX10	CD31	CD3	CD44	CD8	CD57
Mart1	pRB	CD15	CD11b	Ki-67	PDGFRβ
cleavedPARP	MPO	HLA-DR	IL10	Caveolin-1	CD4
CD15	CK5	CXCL12	HLA-ABC	CD31-vWF	CXCL12
CD38	HLA-DR	GranzymeB	Ki67	HLA-ABC	panCK
S100	Cadherin11	HistoneH3	CXCR3	HER2 (D8F12)	
FAP		Galectin9	YAP		
		CD14	CK19		

Table 1: Protein and RNA panels imaged for each of the IMC data sets, with RNA targets bolded

of the three groups is less than 2%, and 3) the training set contains at least 65% of total patches. The actual data splits used in the paper are described in Table 2.

Data set	Group	Patient count	Patch count	Proportion of patches with CD8+ T cells
Metastatic melanoma	Training	102	23741	29.6%
	Validation	28	6045	30.3%
	Testing	29	5950	30.4%
CRC with liver metastases	Training	19	44449	15.9%
	Validation	4	6957	14.4%
	Testing	7	14907	15.9%
Breast cancer	Training	485	41104	23.7%
	Validation	113	9015	23.4%
	Testing	151	12987	23.8%

Table 2: Data split for Melanoma, CRC cohort, and breast tumor IMC data set

Single-cell phenotyping

For each data set, we used the cell type classification (tumor and CD8+ T cells) from the original paper. For the melanoma data set, cell phenotyping was performed using the Shiny application of the R package cytomapper [46], which allows labeling of cell populations using multiple gates. CD8+ T cells were defined using CD3 and CD8, tumor cells are positive for any or multiple of SOX9, SOX10, MITF, Mart1, S100A1, and p75. For the CRC and breast cancer data set, cell type labeling was performed using PhenoGraph [47].

Classifier training

In this work, we trained three classes of models to perform our T cell prediction task. All models presented in this paper were trained with early stopping based on the validation Matthews Correlation Coefficient (MCC) for 10-20 epochs. All models were trained on an NVIDIA GeForce RTX 3090 Ti GPU using PyTorch version 1.13.1 [48]. More details about hyperparameters and implementations can be found in our Github repository.

T cell masking strategy

The purpose of model training is for the model to learn molecular features of the CD8+ T cell's environment that is indicative of its presence, so it is important for us to remove features of the image that are predictive of CD8+ T cell presence but are not part of the cell's environment. We devised a non-trivial cell masking strategy in order to remove T-cell expression patterns without introducing new features that are highly predictive of T cell presence but are not biologically relevant. A simple masking strategy of zeroing out all pixels belonging to CD8+ T cells will introduce contiguous regions of zeros to image patches with T cells, which is an artificial feature that is nonetheless highly predictive of T-cell presence and hence will likely be the main feature learned by a model during training. To circumvent this issue, we first apply a cell "pixelation" step to the original IMC image where we reduce each cell to a single pixel positioned at the cell's centroid. The value of this pixel is the sum of all pixels originally associated with the cell, representing the total signal from each channel within the cell. We then mask this "pixelated" image by zeroing all pixels representing CD8+ T cells. Since there are usually at most two T cell pixels in an image patch, zeroing them in a 16×16 pixel image where most ($> 90\%$) of the pixels are already zeroes is not likely to introduce a significant signal that is predictive T cell presence. We show that our strategy is effective at masking T cells without introducing additional features through a series of image augmentation experiments (Supplemental Note 1 Assessement of T-cell masking strategy).

Logistic regression models

We trained a single-layer neural network on the average intensity values from each molecular channel to obtain a logistic regression classifier, predicting the

probability of CD8+ T cell presence in the image patch. This model represents a linear model where only the average intensity of each molecule is used for prediction instead of their spatial distribution within a patch.

MLP models

Similar to a logistic regression model, the Multilayer Perceptron (MLP) also uses averaged intensity as input features for prediction but is capable of learning nonlinear interactions between features. The MLP model consists of two hidden layers (30 and 10 nodes) with ReLU activation.

U-Net models

To train networks that can make full use of the spatial information, we used a fully convolutional neural network with the U-Net architecture. The U-Net architecture consists of a contracting path and an expansive path, which gives it a U-shaped structure [29]. The contracting path consists of four repeated blocks, each containing a convolutional layer followed by a Rectified Linear Unit (ReLU) activation and a max pooling layer. The expansive path mirrors the contracting path, where each block contains a transposed convolutional layer. Skip connections concatenates the up-sampled features with the corresponding feature maps from the contracting path to include local information. The output of the expansive path is then fed to a fully-connected layer with softmax activation to produce a predicted probability. The model was trained from scratch using image augmentation to prevent over-fitting, including random horizontal/vertical flips and rotations, in addition to standard channel-wise normalization. We train our U-net classifiers using stochastic gradient descent with momentum and a learning rate of 10^{-2} on mini-batches of size 128.

Counterfactual optimization

Given an IMC patch $x^{(i)}$ without T cells, and a classifier f, our goal is to find a perturbation $\delta^{(i)}$ for the patch such that f classifies the perturbed patch as having T cells. For CNN models, $\delta^{(i)} \in \mathbb{R}^{w \times l \times d}$ is a 3D tensor that describes changes made for every channel, at each pixel of the patch.

Given a CNN classifier f and a IMC patch $x^{(i)}$ such that $f(x_0^{(i)}) = \mathbb{P}(\text{T cells present}) < p$, where $p > 0$ is the classification threshold below which the classifier predicts no T-cell, we aim to obtain a perturbation $\delta^{(i)}$ such that $f(x_0^{(i)} + \delta^{(i)}) > p$, by solving the following optimization problem adopted from [30],

$$\delta^{(i)} = \min_{\delta} L_{\text{pred}}(x_0^{(i)}, \delta) + L_{\text{dist}}(\delta) + L_{\text{proto}}(x_0^{(i)}, \delta), \tag{7}$$

such that

$$L_{\text{pred}}(x_0^{(i)}, \delta) = c \max(-f(x_0^{(i)} + \delta), -p), \tag{8}$$

$$L_{\text{dist}}(\delta) = \beta\|\delta\|_1 + \|\delta\|_2^2, \tag{9}$$

$$L_{\text{proto}}(x_0^{(i)}, \delta) = \theta\|x_0^{(i)} + \delta - \text{proto}^{(i)}\|_2^2, \tag{10}$$

$$\delta^{(i)} = \gamma^{(i)} \odot_3 x_0^{(i)} \tag{11}$$

where $\text{proto}^{(i)}$ is an instance of the training set classified as having T cells, defined by first building a k-d tree of training instances classified as having T cells and setting the k-nearest item in the tree (in terms of euclidean distance to $x_0^{(i)}$) as proto. We use $k = 1$ for all counterfactual optimization. For all other parameters, we list their values in Table 3. During optimization, the weight c of the loss term L_{pred} is updated for n iterations, starting at c_0. If we identify a valid counterfactual for the present value of c, we will then decrease c in the subsequent optimization cycle to increase the weight of the additional loss components, thereby enhancing the overall solution. If, however, we do not identify a counterfactual, c is increased to put more emphasis on increasing the predicted probability of the counterfactual. The parameter $s_{\max}$ sets the maximum number of optimization steps for each value of c.

Parameters	Melanoma	CRC
β	2	80
θ	60	40
p	0.5	0.43
c_0	1000	1000
n	5	5
$s_{\max}$	1000	1000

Table 3: Parameter values used for counterfactual optimization

Code Availability

Code for model training, perturbation optimization and analysis are publicly available at https://github.com/neonine2/morpheus. Our optimization code was implemented in Python and was built upon the open source Python library Alibi [49].

Data Availability

All data sets used in this study are published and publicly available.

Acknowledgements

We would like to thank Inna Strazhnik for her support with figure illustrations. We would like to thank Akil Merchant, Alma Andersson, Aviv Regev, Long Cai, Barbara Wold, Michal Polonsky, Jonathan Fox, Yujing Yang, Abdullah Farooq and all members of the Thomson lab for insightful discussion that

significantly improved this work. We gratefully acknowledge the support of the National Institutes of Health's Information Technology for Cancer Research (ITCR) program and the Merkin Institute for Translational Research.

References

[1] Fridman, W. H., Zitvogel, L., Sautès-Fridman, C. & Kroemer, G. The immune contexture in cancer prognosis and treatment. *Nature reviews Clinical oncology* **14** (12), 717–734 (2017) .

[2] Binnewies, M. *et al.* Understanding the tumor immune microenvironment (time) for effective therapy. *Nature medicine* **24** (5), 541–550 (2018) .

[3] Bruni, D., Angell, H. K. & Galon, J. The immune contexture and immunoscore in cancer prognosis and therapeutic efficacy. *Nature Reviews Cancer* **20** (11), 662–680 (2020) .

[4] Hegde, P. S. & Chen, D. S. Top 10 challenges in cancer immunotherapy. *Immunity* **52** (1), 17–35 (2020) .

[5] Choe, J. H., Williams, J. Z. & Lim, W. A. Engineering t cells to treat cancer: the convergence of immuno-oncology and synthetic biology. *Annual Review of Cancer Biology* **4**, 121–139 (2020) .

[6] Pitt, J. *et al.* Targeting the tumor microenvironment: removing obstruction to anticancer immune responses and immunotherapy. *Annals of Oncology* **27** (8), 1482–1492 (2016) .

[7] Haslam, A. & Prasad, V. Estimation of the percentage of us patients with cancer who are eligible for and respond to checkpoint inhibitor immunotherapy drugs. *JAMA network open* **2** (5), e192535–e192535 (2019) .

[8] Lee, J. S. & Ruppin, E. Multiomics Prediction of Response Rates to Therapies to Inhibit Programmed Cell Death 1 and Programmed Cell Death 1 Ligand 1. *JAMA oncology* **5** (11), 1614–1618 (2019) .

[9] Pittet, M. J., Michielin, O. & Migliorini, D. Clinical relevance of tumour-associated macrophages. *Nature reviews Clinical oncology* **19** (6), 402–421 (2022) .

[10] Bonaventura, P. *et al.* Cold tumors: a therapeutic challenge for immunotherapy. *Frontiers in immunology* **10**, 168 (2019) .

[11] Savas, P. *et al.* Clinical relevance of host immunity in breast cancer: from TILs to the clinic. *Nature reviews Clinical oncology* **13** (4), 228–241 (2016) .

[12] Tsaur, I., Brandt, M. P., Juengel, E., Manceau, C. & Ploussard, G. Immunotherapy in prostate cancer: new horizon of hurdles and hopes. *World journal of urology* **39**, 1387–1403 (2021) .

[13] Moffitt, J. R., Lundberg, E. & Heyn, H. The emerging landscape of spatial profiling technologies. *Nature Reviews Genetics* **23** (12), 741–759 (2022) .

[14] Lanitis, E., Dangaj, D., Irving, M. & Coukos, G. Mechanisms regulating t-cell infiltration and activity in solid tumors. *Annals of Oncology* **28**, xii18–xii32 (2017) .

[15] Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363** (6434), 1463–1467 (2019) .

[16] Eng, C.-H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **568** (7751), 235–239 (2019) .

[17] Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature methods* **11** (4), 417–422 (2014) .

[18] Goltsev, Y. *et al.* Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell* **174** (4), 968–981 (2018) .

[19] Bhate, S. S., Barlow, G. L., Schürch, C. M. & Nolan, G. P. Tissue schematics map the specialization of immune tissue motifs and their appropriation by tumors. *Cell Systems* **13** (2), 109–130 (2022) .

[20] Wu, Z. *et al.* Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens. *Nature Biomedical Engineering* 1–14 (2022) .

[21] Schürch, C. M. *et al.* Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front. *Cell* **182** (5), 1341–1359 (2020) .

[22] Aoki, T. *et al.* The spatially resolved tumor microenvironment predicts treatment outcome in relapsed/refractory Hodgkin lymphoma. *bioRxiv* 2023–05 (2023) .

[23] Chang, C.-H., Creager, E., Goldenberg, A. & Duvenaud, D. Explaining Image Classifiers by Counterfactual Generation. *International Conference on Learning Representations (ICLR)* (2019) .

[24] Wachter, S., Mittelstadt, B. & Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* **31**, 841 (2017) .

[25] Verma, S. *et al.* Counterfactual explanations and algorithmic recourses for machine learning: A review. *arXiv preprint arXiv:2010.10596* (2020) .

[26] Hoch, T. *et al.* Multiplexed imaging mass cytometry of the chemokine milieus in melanoma characterizes features of the response to immunotherapy. *Science Immunology* **7** (70) (2022) .

[27] Wang, Z. *et al.* Extracellular vesicles in fatty liver promote a metastatic tumor microenvironment. *Cell Metabolism* (2023) .

[28] Danenberg, E. *et al.* Breast tumor microenvironment structures are associated with genomic features and clinical outcome. *Nature genetics* **54** (5), 660–669 (2022) .

[29] Buda, M., Saha, A. & Mazurowski, M. A. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in Biology and Medicine* **109** (2019). https://doi.org/10.1016/j.compbiomed.2019.05.002 .

[30] Looveren, A. V. & Klaise, J. Interpretable counterfactual explanations guided by prototypes. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 650–665 (2021) .

[31] Hughes, C. E. & Nibbs, R. J. A guide to chemokines and their receptors. *The FEBS journal* **285** (16), 2944–2971 (2018) .

[32] Foeng, J., Comerford, I. & McColl, S. R. Harnessing the chemokine system to home CAR-T cells into solid tumors. *Cell Reports Medicine* (2022) .

[33] Kohli, K., Pillarisetty, V. G. & Kim, T. S. Key chemokines direct migration of immune cells in solid tumors. *Cancer gene therapy* **29** (1), 10–21 (2022) .

[34] Schraufstatter, I. U., Zhao, M., Khaldoyanidi, S. K. & DiScipio, R. G. The chemokine CCL18 causes maturation of cultured monocytes to macrophages in the M2 spectrum. *Immunology* **135** (4), 287–298 (2012) .

[35] McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints* (2018). https://arxiv.org/abs/1802.03426 .

[36] Ghanem, I. *et al.* Insights on the CXCL12-CXCR4 axis in hepatocellular carcinoma carcinogenesis. *American journal of translational research* **6** (4), 340 (2014) .

[37] Biasci, D. *et al.* CXCR4 inhibition in human pancreatic and colorectal cancers induces an integrated immune response. *Proceedings of the National Academy of Sciences* **117** (46), 28960–28970 (2020) .

[38] Chen, Y. *et al.* CXCR4 inhibition in tumor microenvironment facilitates anti-programmed death receptor-1 immunotherapy in sorafenib-treated hepatocellular carcinoma in mice. *Hepatology* **61** (5), 1591–1602 (2015) .

[39] Steele, M. M. *et al.* T cell egress via lymphatic vessels is tuned by antigen encounter and limits tumor control. *Nature Immunology* **24** (4), 664–675 (2023) .

[40] Sullivan, K. M. *et al.* Blockade of interleukin 10 potentiates antitumour immune function in human colorectal cancer liver metastases. *Gut* **72** (2), 325–337 (2023) .

[41] Zhang, H., Pasolli, H. A. & Fuchs, E. Yes-associated protein (YAP) transcriptional coactivator functions in balancing growth and differentiation in skin. *Proceedings of the National Academy of Sciences* **108** (6), 2270–2275 (2011) .

[42] Chen, A. *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using dna nanoball-patterned arrays. *Cell* **185** (10), 1777–1792 (2022) .

[43] Janesick, A. *et al.* High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of FFPE tissue. *Biorxiv* 2022–10 (2022) .

[44] Qian, J. *et al.* The IFN-γ/PD-L1 axis between T cells and tumor microenvironment: hints for glioma anti-PD-1/PD-L1 therapy. *Journal of neuroinflammation* **15** (1), 1–13 (2018) .

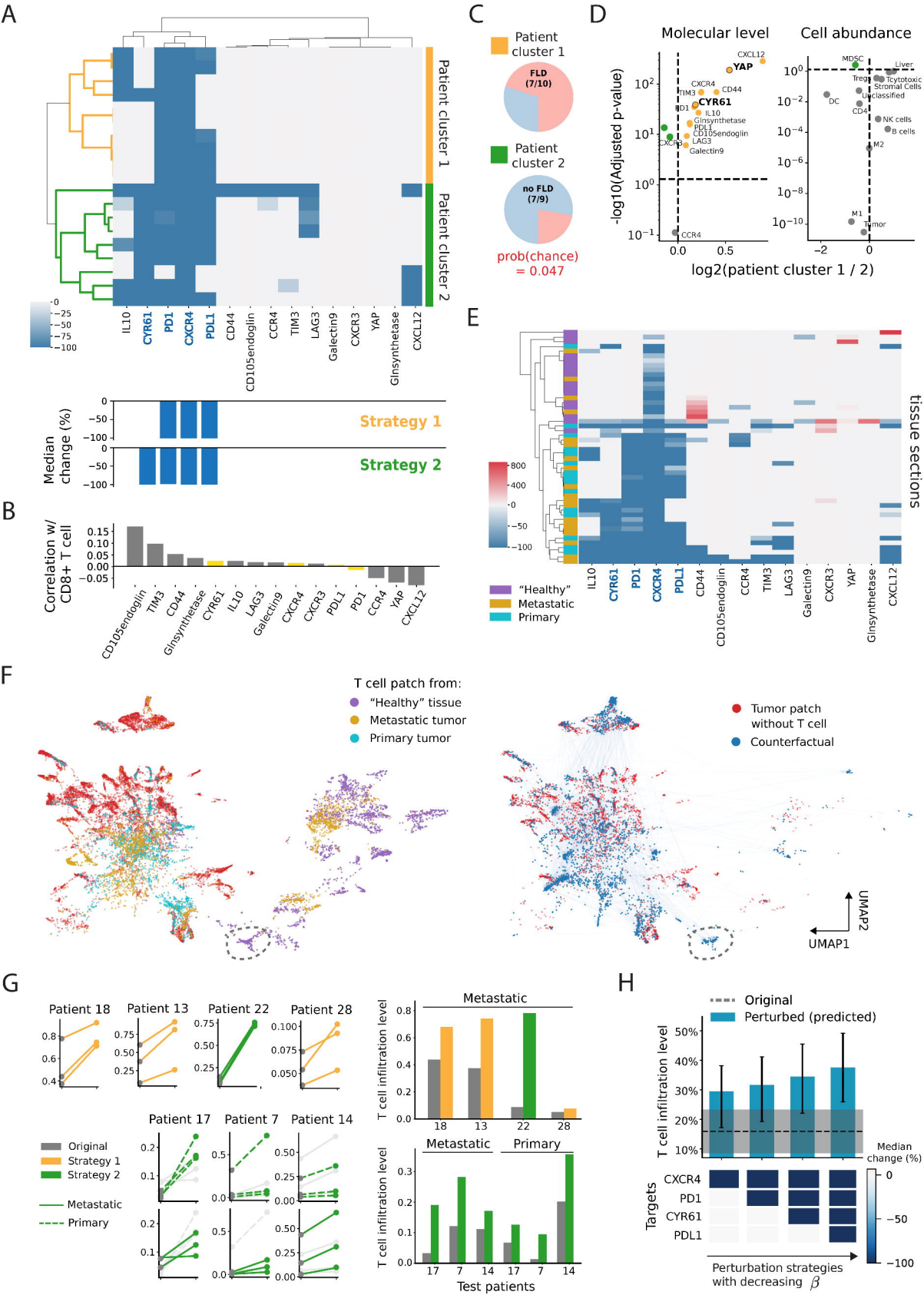
[45] Consortium, G. O. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research* **32**, D258–D261 (2004) .

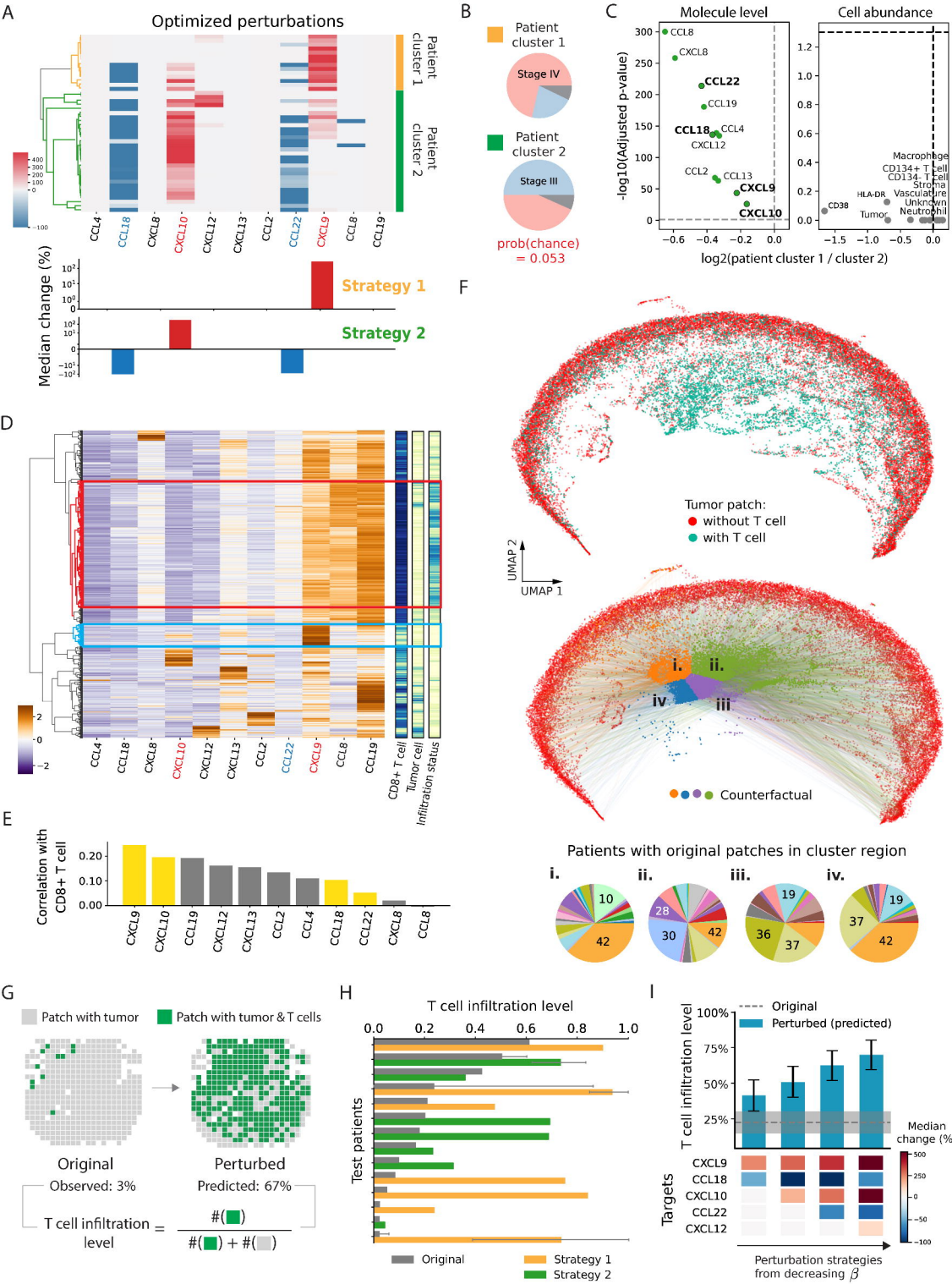
[46] Eling, N., Damond, N., Hoch, T. & Bodenmiller, B. cytomapper: an R/Bioconductor package for visualization of highly multiplexed imaging data. *Bioinformatics* **36** (24), 5706–5708 (2020) .

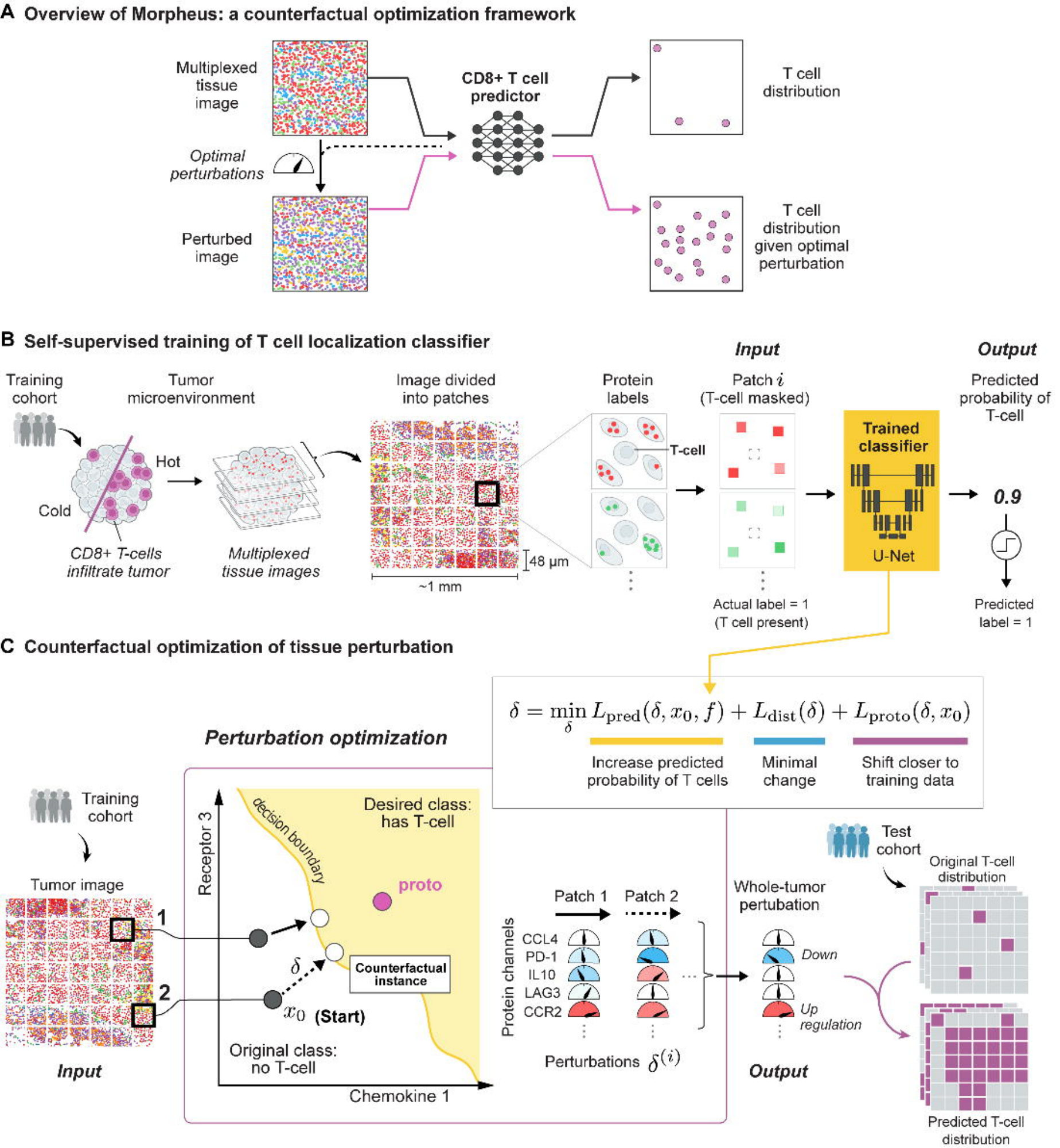
[47] Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162** (1), 184–197 (2015) .

[48] Paszke, A. *et al.* PyTorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019) .

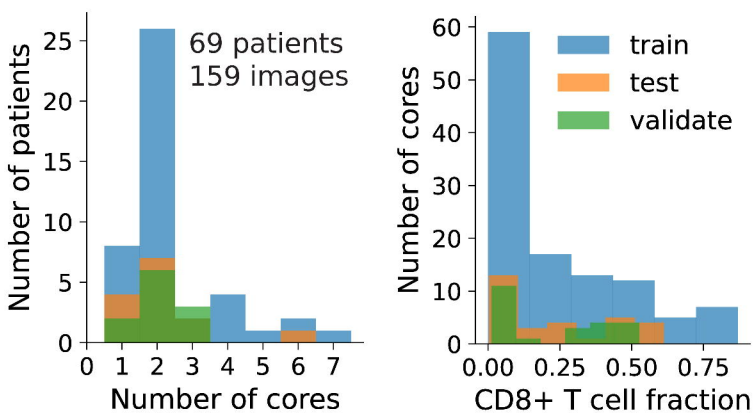
[49] Klaise, J., Looveren, A. V., Vacanti, G. & Coca, A. Alibi Explain: Algorithms for Explaining Machine Learning Models. *Journal of Machine Learning Research* **22** (181), 1–7 (2021) .

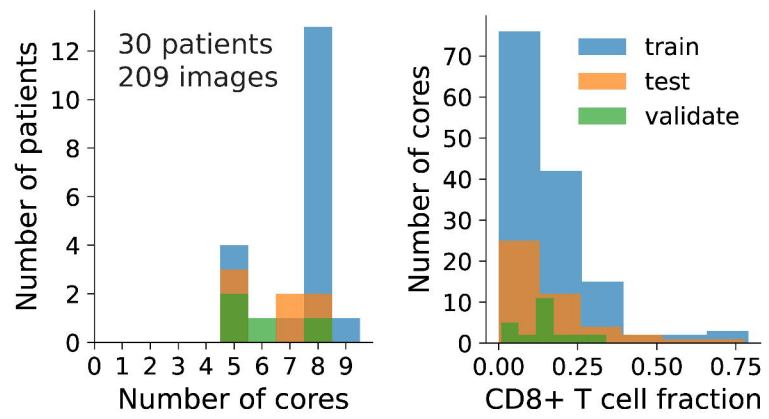

A
Patient cluster 1
Patient cluster 2
IL10
CYR61
PD1
CXCR4
PDL1
CD44
CD105endoglin
CCR4
TIM3
LAG3
Galectin9
CXCR3
YAP
Glnsynthetase
CXCL12
Median change (%)
Strategy 1
Strategy 2
B
Correlation w/ CD8+ T cell
C
Patient cluster 1
FLD (7/10)
Patient cluster 2
no FLD (7/9)
prob(chance) = 0.047
D
Molecular level
Cell abundance
-log10(Adjusted p-value)
log2(patient cluster 1 / 2)
E
tissue sections
"Healthy"
Metastatic
Primary
F
T cell patch from:
"Healthy" tissue
Metastatic tumor
Primary tumor
Tumor patch without T cell
Counterfactual
UMAP1
UMAP2
G
Patient 18
Patient 13
Patient 22
Patient 28
Patient 17
Patient 7
Patient 14
Original
Strategy 1
Strategy 2
Metastatic
Primary
T cell infiltration level
Test patients
H
Original
Perturbed (predicted)
T cell infiltration level
Median change (%)
Targets
CXCR4
PD1
CYR61
PDL1
Perturbation strategies with decreasing β

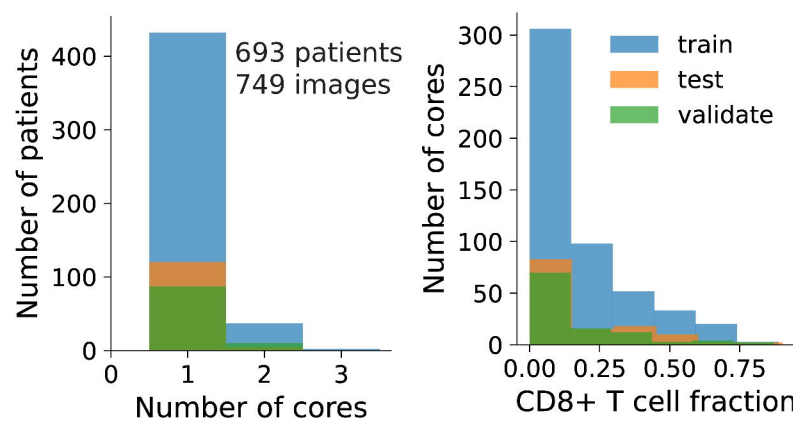

A
Optimized perturbations
Patient cluster 1
Patient cluster 2
CCL4
CCL18
CXCL8
CXCL10
CXCL12
CXCL13
CCL2
CCL22
CXCL9
CCL8
CCL19
Median change (%)
Strategy 1
Strategy 2
B
Patient cluster 1
Stage IV
Patient cluster 2
Stage III
prob(chance) = 0.053
C
Molecule level
Cell abundance
-log10(Adjusted p-value)
log2(patient cluster 1 / cluster 2)
CCL8
CXCL8
CCL22
CCL19
CCL18
CCL4
CXCL12
CCL2
CCL13
CXCL9
CXCL10
Macrophage
CD134+ T cell
CD134- T cell
Stroma
Vasculature
Unknown
Neutrophil
HLA-DR
CD38
Tumor
D
CD8+ T cell
Tumor cell
Infiltration status
E
Correlation with CD8+ T cell
CXCL9
CXCL10
CCL19
CXCL12
CXCL13
CCL2
CCL4
CCL18
CCL22
CXCL8
CCL8
F
Tumor patch:
without T cell
with T cell
UMAP 2
UMAP 1
i.
ii.
iii.
iv
Counterfactual
Patients with original patches in cluster region
i.
ii.
iii.
iv.
10
42
28
30
42
19
36
37
19
37
42
G
Patch with tumor
Patch with tumor & T cells
Original
Observed: 3%
Perturbed
Predicted: 67%
T cell infiltration level = #(■) / #(■) + #(□)
H
T cell infiltration level
Test patients
Original
Strategy 1
Strategy 2
I
Original
Perturbed (predicted)
T cell infiltration level
Median change (%
Targets
CXCL9
CCL18
CXCL10
CCL22
CXCL12
Perturbation strategies from decreasing β


A Overview of Morpheus: a counterfactual optimization framework
Multiplexed tissue image
CD8+ T cell predictor
T cell distribution
Optimal perturbations
Perturbed image
T cell distribution given optimal perturbation
B Self-supervised training of T cell localization classifier
Input
Output
Training cohort
Tumor microenvironment
Image divided into patches
Protein labels
Patch i (T-cell masked)
Predicted probability of T-cell
Hot
Cold
T-cell
Trained classifier
0.9
CD8+ T-cells infiltrate tumor
Multiplexed tissue images
48 µm
~1 mm
U-Net
Actual label = 1 (T cell present)
Predicted label = 1
C Counterfactual optimization of tissue perturbation
$\delta = \min_{\delta} L_{\text{pred}}(\delta, x_0, f) + L_{\text{dist}}(\delta) + L_{\text{proto}}(\delta, x_0)$
Increase predicted probability of T cells
Minimal change
Shift closer to training data
Perturbation optimization
Training cohort
Tumor image
Input
Receptor 3
decision boundary
Desired class: has T-cell
proto
Counterfactual instance
δ
x_0 (Start)
Original class: no T-cell
Chemokine 1
Patch 1
Patch 2
Protein channels
CCL4
PD-1
IL10
LAG3
CCR2
Perturbations $\delta^{(i)}$
Test cohort
Whole-tumor pertubation
Down
Up regulation
Output
Original T-cell distribution
Predicted T-cell distribution

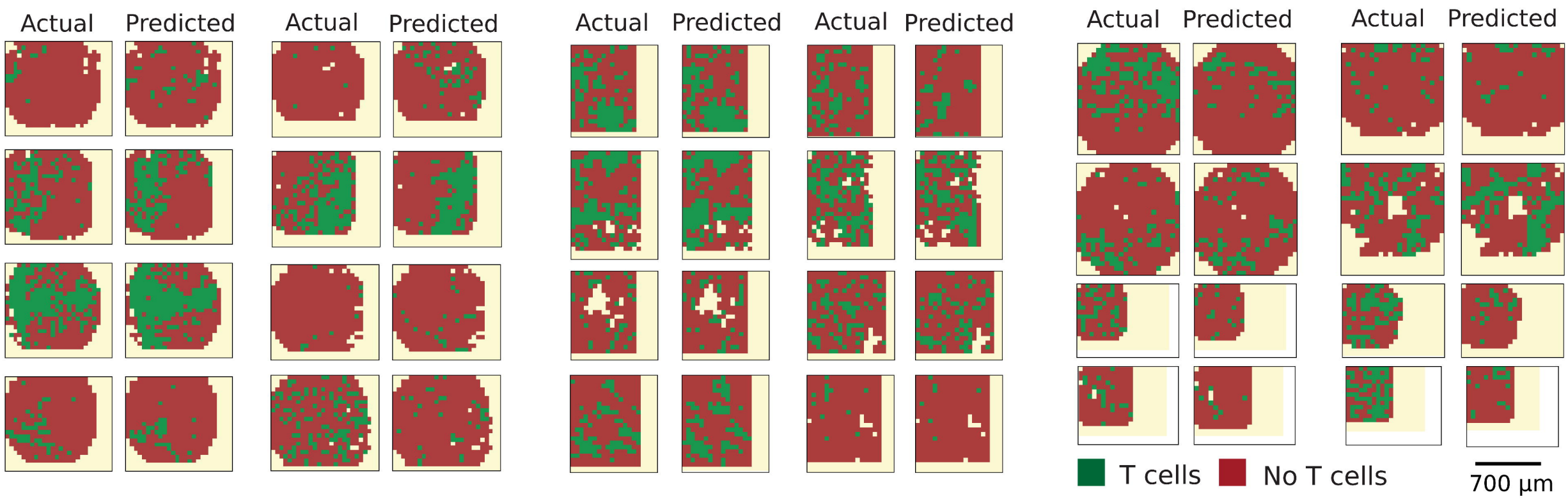
A


Melanoma
69 patients
159 images
Number of patients
Number of cores
Number of cores
CD8+ T cell fraction
train
test
validate

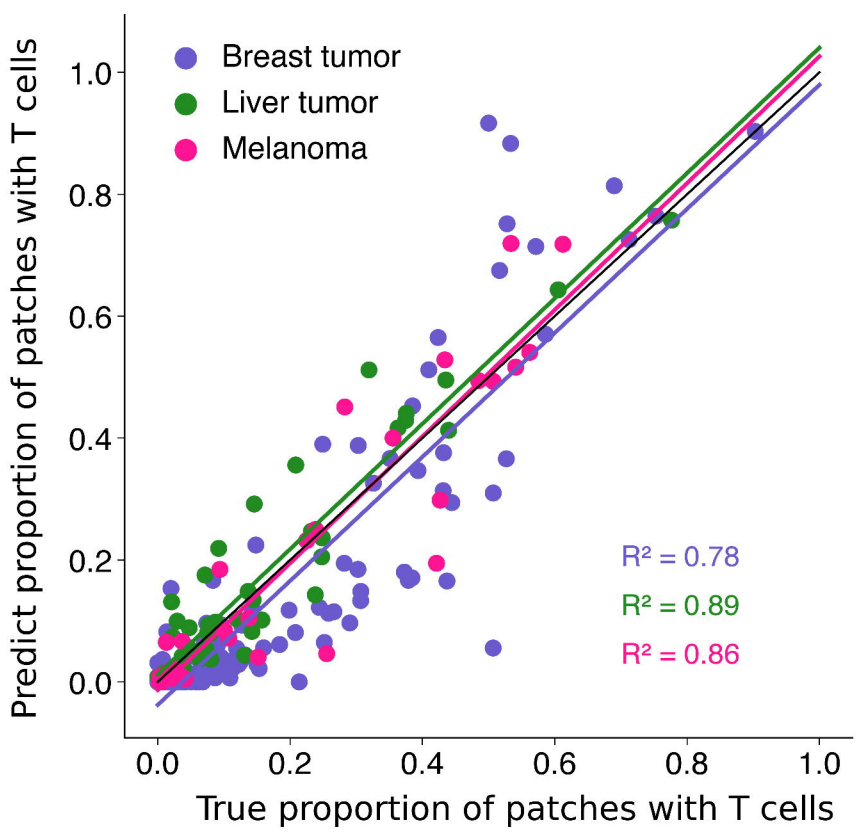

CRC with liver metastases
30 patients
209 images
Number of patients
Number of cores
Number of cores
CD8+ T cell fraction
train
test
validate


Breast tumor
693 patients
749 images
Number of patients
Number of cores
Number of cores
CD8+ T cell fraction
train
test
validate

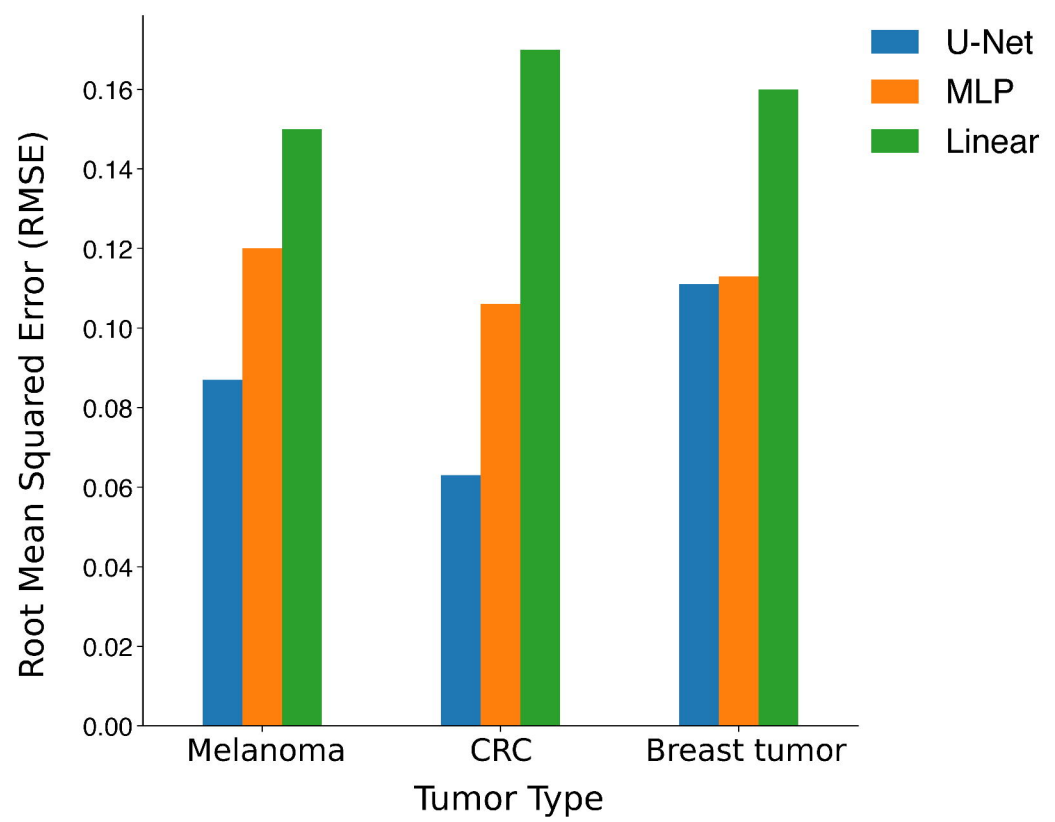
B


Melanoma
Actual
Predicted
Actual
Predicted
CRC with liver metastases
Actual
Predicted
Actual
Predicted
Breast tumor
Actual
Predicted
Actual
Predicted
T cells
No T cells
700 µm

C


Breast tumor
Liver tumor
Melanoma
R² = 0.78
R² = 0.89
R² = 0.86
Predict proportion of patches with T cells
True proportion of patches with T cells

D


U-Net
MLP
Linear
Root Mean Squared Error (RMSE)
Melanoma
CRC
Breast tumor
Tumor Type